Subspace Models for Acoustic Unit Discovery

Lucas Ondel, Bolaji Yusuf

Brno University of Technology, Faculty of Information Technology
Božetěchova 1/2. 612 66 Brno - Královo Pole
{iondel,xyusuf00}@fit.vutbr.cz

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

January 26, 2021

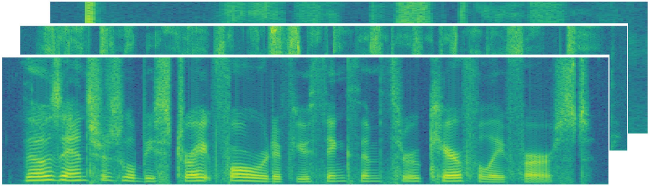
Table of Contents

Outline

- **Acoustic Unit Discovery**
  - Definition of the task
  - Applications
- **State-of-the-art**
  - Major approaches
  - Evaluations
- **Subspace Models For AUD**
  - Motivations
  - Subspace HMM
  - Hierarchical Subspace HMM
- **Conclusion**

Definition of the task

- Audio recordings without labels
- Inventory of phone-like units (we call them “acoustic units”)
- Segmentation and labelling



Learning like a baby

Outline

- **Acoustic Unit Discovery**
  - Definition of the task
  - Applications
- **State-of-the-art**
  - Major approaches
  - Evaluations
- **Subspace Models For AUD**
  - Motivations
  - Subspace HMM
  - Hierarchical Subspace HMM
- **Conclusion**

Documenting endangered languages

- Language diversity is diminishing world wide
  - Speech technologies are only available for a few languages
  - Majority of languages are not written: main-stream ASR is not applicable
- an accurate AUD system could:
  - help linguists to document languages
  - serve as a front-end of speech-technologies for non-written languages
- 2022-2032: The decade of Indigenous Languages

UNESCO website

- The learning cognitive process of humans remains largely unknown:
  - the brain is very complex
  - "sensory learning" happens at a very early age, when the infants cannot communicate verbally
- Reverse engineering approach ("E. Dupoux. 2018") Link:
  - let's build model to learn speech in an unsupervised fashion
  - analyse it and derive the learning principle
  - pave the way to a more realistic artificial intelligence

Another type of machine learning

- The "always more data" approach raises concerns:
  - social/ethical problems: monopoly of ML technologies by LARGE data owner
  - ecological issues: more data implies more energy consumption
- AUD research implies data efficient models

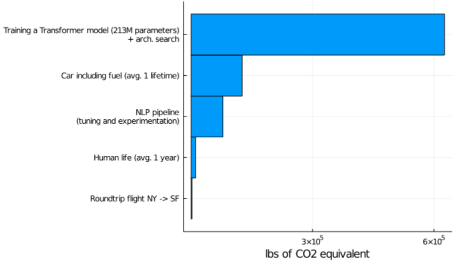


Figure: Energy consumption of training deep learning models. Source: Strubell *et al*, 2019 Link

Outline

Outline

FIT

- heuristic-based model: $\sim$ 1990 – 2005
- non-parametric Bayesian-based models: $\sim$ 2005 – 2020
- neural network-based models: $\sim$ 2015 – 2020

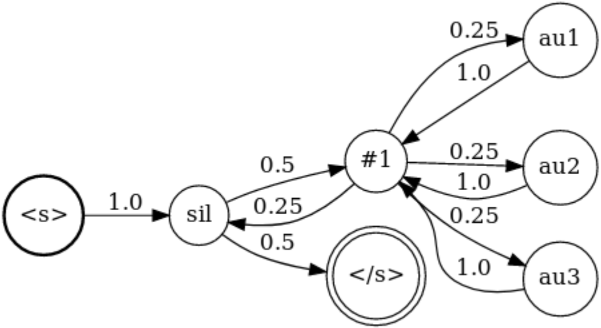
Neural network-based models

- VAE-HMM: Variational AutoEncoder with HMM/GMM prior over the latent space. Link
- Visually guided neural network: replace textual transcription by images Link
- VQ-VAE: AutoEncoder with discretized bottleneck Link

Non-parametric Bayesian-based models

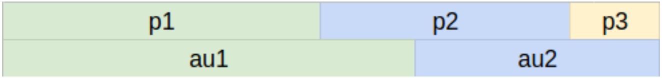
- Non-parametric model for word segmentation Link 1 Link 2
- Dirichlet-Process HMM (2012) Link
- Dirichlet-Process HMM (Variational Bayes inference) (2016) Link

Dirichlet-Process HMM baseline



Outline

Metrics



- Clustering:
  - Normalized Mutual Information: $200 \frac{MI(X,Y)}{H(X)+H(Y)}\%$
  - Measures the statistical relation between the discovered units and the ground truth transcription
  - 100 % $\rightarrow$ one-to-one mapping between AU and phones
  - 0 % $\rightarrow$ AUs are completely uninformative
- Segmentation:
  - F-score: harmonic mean of segmentation precision and recall
  - $\pm$ 20 ms tolerance allowed

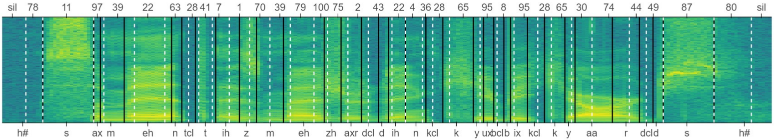
Experiments

- Data:
  - Mboshi (Congo Brazzaville) - 3-4 hours
  - Yoruba (West Africa - Nigeria) - 10 hours
  - English (TIMIT) - 4 hours
- Features:
  - traditional features: MFCCs + Δ + $\Delta\Delta$

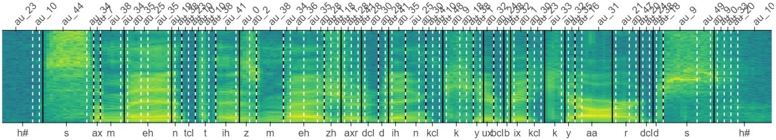
- Adapted version of "Chorowski et al., 2019" Link
- Dirichlet Process HMM (VB inference) Link

Corpus	System	NMI	F-Score	# units
English	VQ-VAE	32.03	59.05	50
English	HMM	**35.91**	**63.86**	95
Mboshi	VQ-VAE	31.27	39.19	50
Mboshi	HMM	**35.87**	**47.92**	94
Yoruba	VQ-VAE	29.90	37.52	50
Yoruba	HMM	**36.38**	**54.47**	95

Table: Comparison of the HMM vs the VQ-VAE baseline


sil 78 11 97 39 22 63 28 41 7 1 70 39 79 100 75 2 43 22 4 36 28 65 95 8 95 28 65 30 74 44 49 87 80 sil
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#

HMM

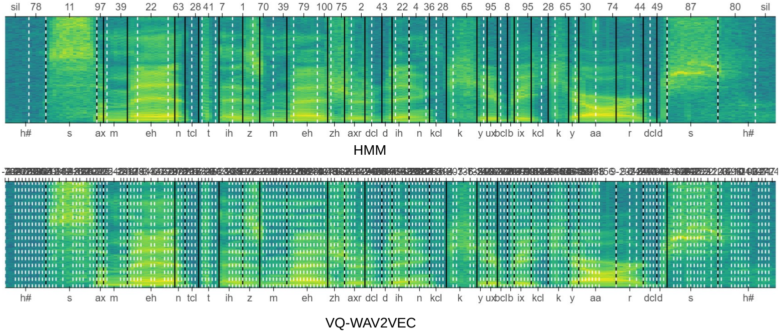

h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#

VQ-VAE

- VQ-WAV2VEC "A. Baevski et al., 2020" trained on 960h of LibriSpeech (unsupervised) Link
- Dirichlet Process HMM (VB inference) Link

Corpus	System	NMI	F-Score	# units
English	VQ-WAV2VEC (Gumbel)	35.20	26.84	12008
English	VQ-WAV2VEC (K-mean)	34.06	25.64	20057
English	HMM	**35.47**	**63.86**	95

Table: Comparison of the HMM vs the VQ-VAE baseline

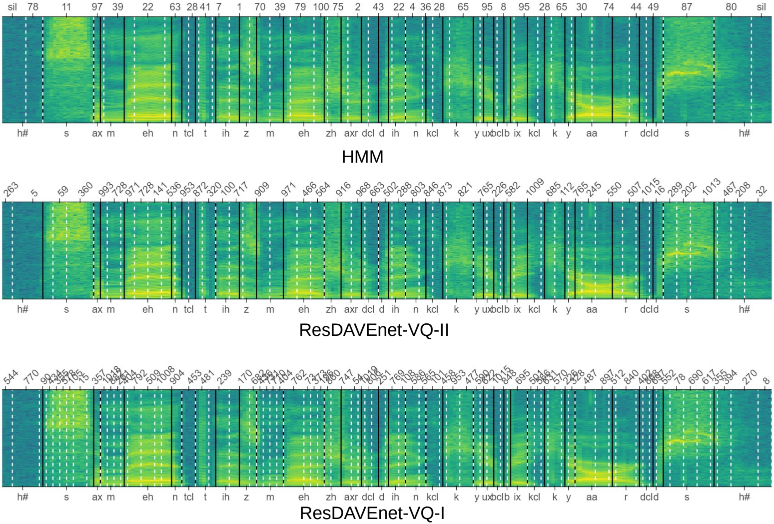

sil 78 11 97 39 22 63 28 41 7 1 70 39 79 100 75 2 43 22 4 36 28 65 95 8 95 28 65 30 74 44 49 87 80 sil
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
HMM
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
VQ-WAV2VEC

- ResDAVEnet-VQ-(I/II) "Harwath et al., 2019" Link
- Dirichlet Process HMM (VB inference) Link

Corpus	System	NMI	F-Score	# units
English	ResDAVEnet-VQ-I	**35.93**	54.19	979
English	ResDAVEnet-VQ-II	34.39	**64.36**	224
English	HMM	35.91	63.86	95

Table: Comparison of the HMM vs the VQ-VAE baseline

Example



sil 78 11 97 39 22 63 28 41 7 1 70 39 79 100 75 2 43 22 4 36 28 65 95 8 95 28 65 30 74 44 49 87 80 sil
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
HMM
263 5 59 360 993 728 971 728 141 536 953 872 320 100 717 909 971 466 564 916 968 663 502 288 803 846 873 821 765 226 582 1009 685 112 765 245 550 507 1015 16 289 202 1013 467 208 32
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
ResDAVEnet-VQ-II
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
ResDAVEnet-VQ-I

Outline

Outline

- **Acoustic Unit Discovery**
  - Definition of the task
  - Applications
- **State-of-the-art**
  - Major approaches
  - Evaluations
- **Subspace Models For AUD**
  - Motivations
  - Subspace HMM
  - Hierarchical Subspace HMM
- **Conclusion**

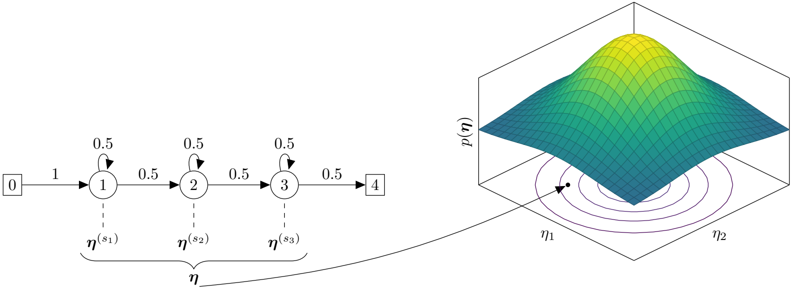
Motivations

- Infants do not learn from scratch (Kuhl *et al*, 1992 Link):
  - They have some innate sensitivity to human languages
  - With time, they become specialized to their native language
- Hypothesis/Design choice:
  - This innate sensitivity guide infants to learn the structure of speech
  - The AUD system should adapt and become language specific
- Proposal: we will use Bayesian Subspace Model techniques to implement these properties:
  - Subspace Hidden Markov Model (SHMM)
  - Hierarchical Subspace Hidden Markov Model (HSHMM)

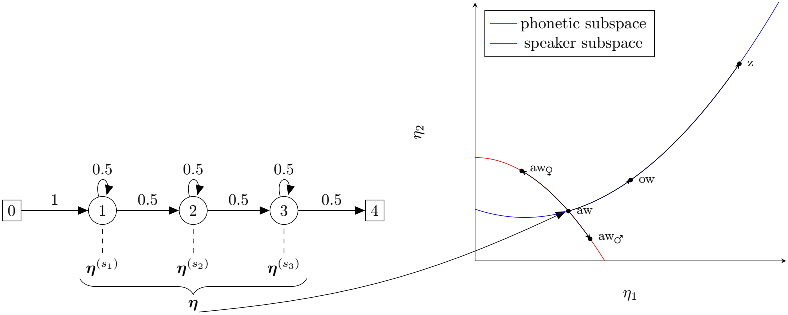
Outline

Learning an acoustic unit

$$p(\boldsymbol{\eta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\eta})}{p(\mathbf{X})} \tag{1}$$



Phonetic Subspace



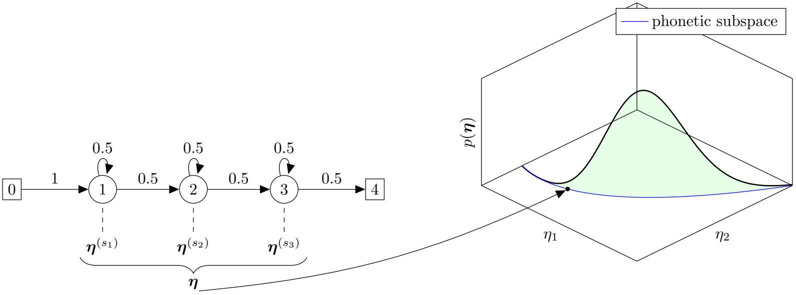
Prior over the the AU's parameters

- We want to design an educated prior over the AU's parameters : $p(\boldsymbol{\eta})$
-

$$\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

$$\boldsymbol{\eta} = f(\mathbf{W}\mathbf{h} + \mathbf{b}) \tag{3}$$

Subspace HMM



- We estimate the subspace parameters $\mathbf{W}, \mathbf{b}$ on annotated corpora
- Model is trained by optimizing the Evidence Lower-BOund (ELBO):

$$
\begin{aligned}
\mathcal{L} = \langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{W}, \mathbf{b}, \mathbf{h}_{1:T}) \rangle_q \\
- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) || p(\mathbf{z})) \\
- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{W}) q(\mathbf{b})) || p(\mathbf{W}) p(\mathbf{b})) \\
- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T}) || p(\mathbf{h}_{1:T}))
\end{aligned}
$$

- The training follows an Expectation-Maximization-like training:
  - E-step: Baum-Welch algorithm to estimate states' occupancy
  - M-step: No closed form solution, using re-parameterization trick.

- The subspace parameters **W**, **b** are fixed, we just learn the embeddings **h** on the target language
- Model is trained by optimizing the Evidence Lower-BOund (ELBO):

$$\begin{aligned}\mathcal{L} = &\langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{W}, \mathbf{b}, \mathbf{h}_{1:T}) \rangle_q \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}))||p(\mathbf{z})) \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T})||p(\mathbf{h}_{1:T}))\end{aligned}$$

- The training follows an Expectation-Maximization-like training:
  - E-step: Baum-Welch algorithm to estimate states' occupancy
  - M-step: No closed form solution, using re-parameterization trick.

Illustration

Link

Experiments

- Data:
  - Source languages (transcribed)
    - French, German, Polish, Spanish from Globalphone
    - 3-4 hours subsets of each language's training data
  - Target languages (untranscribed)
    - Mboshi (Congo Brazzaville) - 3-4 hours
    - Yoruba (West Africa - Nigeria) - 10 hours
    - English (TIMIT) - 4 hours
- Features:
  - traditional features: MFCCs + Δ + $\Delta\Delta$

Corpus	System	Training	NMI	F-Score
English	HMM	no	1.74	0.20
English	SHMM	no	20.83	58.94
Mboshi	HMM	no	1.65	0.02
Mboshi	SHMM	no	21.0	39.28
Yoruba	HMM	no	1.39	0.45
Yoruba	SHMM	no	22.67	45.83

Table: Comparison of the HMM vs the SHMM before training

Corpus	System	Training	NMI	F-Score
English	HMM	no	1.74	0.20
English	HMM	yes	35.91	63.86
English	SHMM	no	20.83	58.94
English	SHMM	yes	**39.17**	**74.65**
Mboshi	HMM	no	1.65	0.02
Mboshi	HMM	yes	35.85	47.92
Mboshi	SHMM	no	21.0	39.28
Mboshi	SHMM	yes	**38.38**	**59.50**
Yoruba	HMM	no	1.39	0.45
Yoruba	HMM	yes	36.38	54.47
Yoruba	SHMM	no	22.67	45.83
Yoruba	SHMM	yes	**38.99**	**64.46**

Table: Comparison of the HMM vs the SHMM before training

Corpus	System	Training	NMI	F-Score
English	HMM	no	1.74	0.20
English	HMM	yes	35.91	63.86
English	SHMM	no	20.83	58.94
English	SHMM	yes	**39.17**	**74.65**
English	SHMM (2)	yes	37.83	72.20
Mboshi	HMM	no	1.65	0.02
Mboshi	HMM	yes	35.85	47.92
Mboshi	SHMM	no	21.0	39.28
Mboshi	SHMM	yes	**38.38**	**59.50**
Mboshi	SHMM (2)	yes	36.09	53.06
Yoruba	HMM	no	1.39	0.45
Yoruba	HMM	yes	36.38	54.47
Yoruba	SHMM	no	22.67	45.83
Yoruba	SHMM	yes	**38.99**	**64.46**
Yoruba	SHMM (2)	yes	36.97	58.59

Table: Comparison of the HMM, SHMM and HSHMM

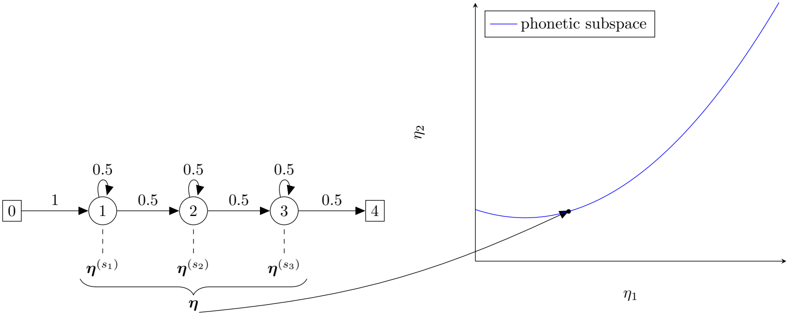
SHMM (2): the subspace is retrained on the target language.

Outline

- **Acoustic Unit Discovery**
  - Definition of the task
  - Applications
- **State-of-the-art**
  - Major approaches
  - Evaluations
- **Subspace Models For AUD**
  - Motivations
  - Subspace HMM
  - Hierarchical Subspace HMM
- **Conclusion**

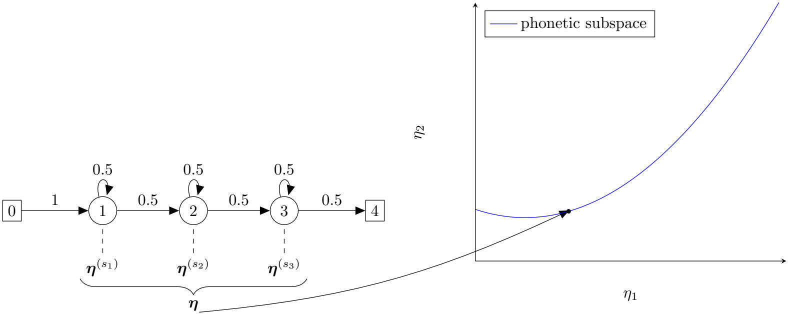
Revisiting the Phonetic Subspace

- We assume the subspace is known and fixed during AUD
- Subspace is the same for all the target languages

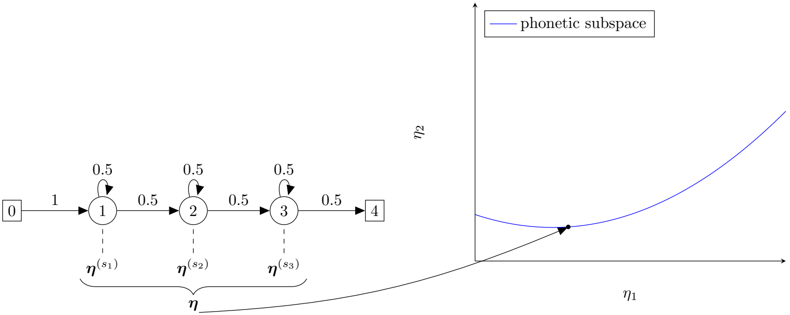


Revisiting the Phonetic Subspace

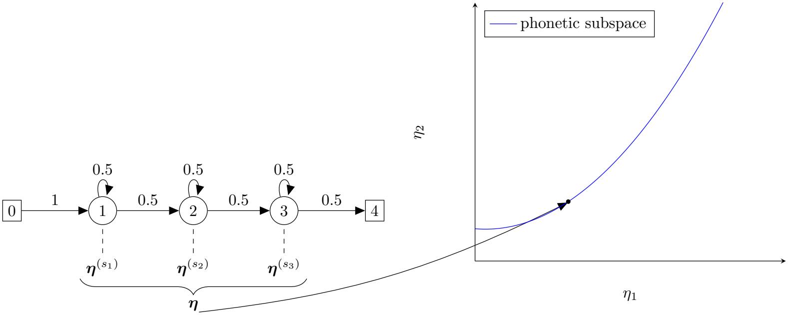
- We would like to adapt the subspace to make it language specific



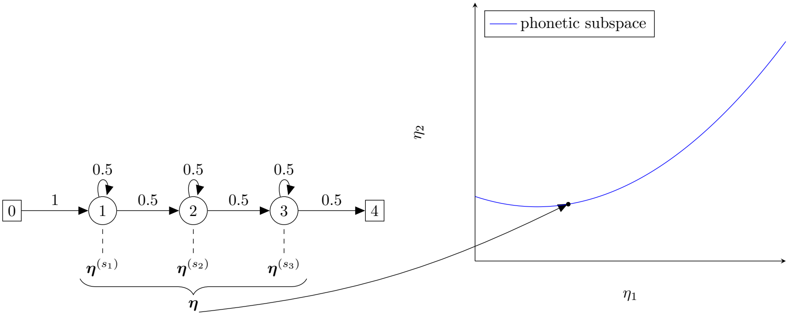
- We would like to adapt the subspace to make it language specific



- We would like to adapt the subspace to make it language specific



- We would like to adapt the subspace to make it language specific



- We want to design an educated prior over all possible subspace: $p(\mathbf{W}, \mathbf{b})$
-

$$\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

$$\mathbf{W} = \sum_{i=1}^{Q} \alpha_i \mathbf{M}_i \tag{5}$$

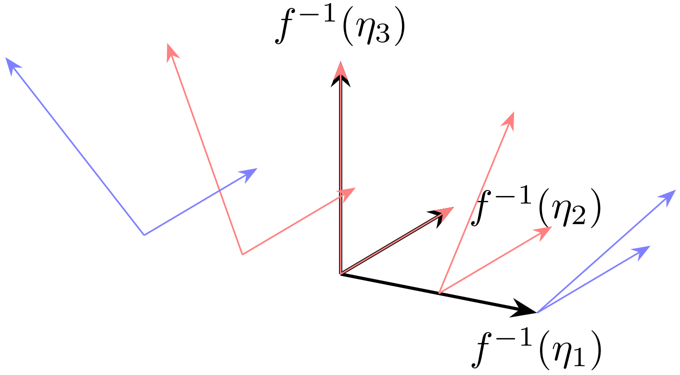
$$\mathbf{b} = \sum_{i=1}^{Q} \alpha_i \mathbf{m}_i \tag{6}$$

$$\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{7}$$

$$\boldsymbol{\eta} = f(\mathbf{W}\mathbf{h} + \mathbf{b}) \tag{8}$$

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^\top \tag{9}$$

$$\mathbf{W} = \alpha_1 \mathbf{M}_1 + \alpha_2 \mathbf{M}_2 \tag{10}$$



- We estimate the "hyper-subspace" parameters $\mathbf{M}, \mathbf{m}, \boldsymbol{\alpha}$ on annotated corpora
- Model is trained by optimizing the Evidence Lower-BOund (ELBO):

$$\begin{aligned}\mathcal{L} = & \langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{M}_{1:Q}, \mathbf{m}_{1:Q}, \mathbf{h}_{1:T}, \boldsymbol{\alpha}) \rangle_q \\ & - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) || p(\mathbf{z})) \\ & - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{M}_{1:Q}) q(\mathbf{m}_{1:Q}) || p(\mathbf{M}_{1:Q}) p(\mathbf{m}_{1:Q})) \\ & - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T}) || p(\mathbf{h}_{1:T})) \\ & - \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\alpha}) || p(\boldsymbol{\alpha}))\end{aligned}$$

- The training follows an Expectation-Maximization-like training:
  - E-step: Baum-Welch algorithm to estimate states' occupancy
  - M-step: No closed form solution, using re-parameterization trick.

- The "hyper-subspace" parameters $\mathbf{M}, \mathbf{m}$ are fixed, we just learn the embeddings $\mathbf{h}$ and α on the target language
- Model is trained by optimizing the Evidence Lower-BOund (ELBO):

$$
\begin{aligned}
\mathcal{L} = \langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{M}_{1:Q}, \mathbf{m}_{1:Q}, \mathbf{h}_{1:T}, \boldsymbol{\alpha}) \rangle_q \\
- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) || p(\mathbf{z})) \\
- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T}) || p(\mathbf{h}_{1:T})) \\
- \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\alpha}) || p(\boldsymbol{\alpha}))
\end{aligned}
$$

- The training follows an Expectation-Maximization-like training:
  - E-step: Baum-Welch algorithm to estimate states' occupancy
  - M-step: No closed form solution, using re-parameterization trick.

Illustration

Link

Experiments

- Data:
  - Source languages (transcribed)
    - French, German, Polish, Spanish from Globalphone
    - 3-4 hours subsets of each language's training data
  - Target languages (untranscribed)
    - Mboshi (Congo Brazzaville) - 3-4 hours
    - Yoruba (West Africa - Nigeria) - 10 hours
    - English (TIMIT) - 4 hours
- Features:
  - traditional features: MFCCs + Δ + $\Delta\Delta$

Results

Corpus	System	NMI	F-Score
English	HMM	35.91	63.86
English	SHMM	39.17	74.65
Mboshi	HMM	35.85	47.92
Mboshi	SHMM	38.38	59.50
Yoruba	HMM	36.38	54.47
Yoruba	SHMM	38.99	64.46

Table: Comparison of the HMM, SHMM and HSHMM

Corpus	System	NMI	F-Score
English	HMM	35.91	63.86
English	SHMM	39.17	74.65
English	SHMM (2)	37.83	72.2
Mboshi	HMM	35.85	47.92
Mboshi	SHMM	38.38	59.50
Mboshi	SHMM (2)	36.09	53.06
Yoruba	HMM	36.38	54.47
Yoruba	SHMM	38.99	64.46
Yoruba	SHMM (2)	36.97	58.59

Table: Comparison of the HMM, SHMM and HSHMM

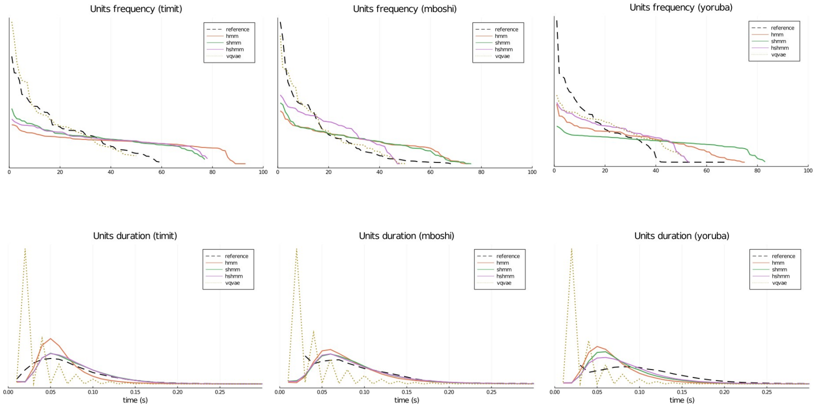
SHMM (2): the subspace is retrained on the target language.

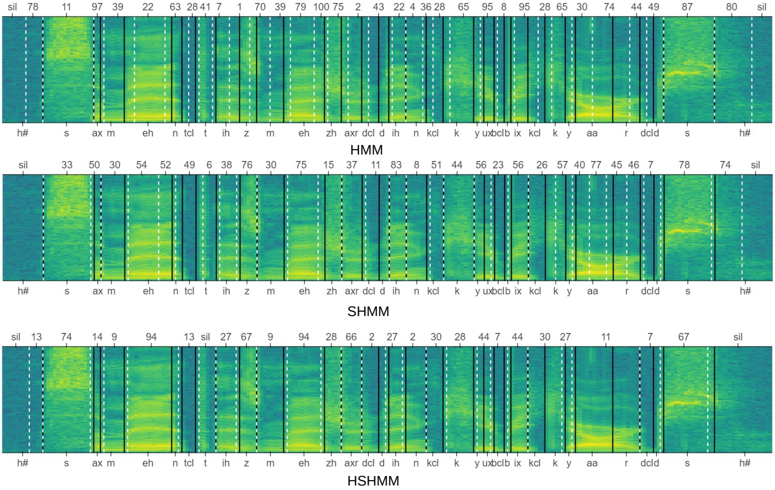
Results

Corpus	System	NMI	F-Score
English	HMM	35.91	63.86
English	SHMM	39.17	74.65
English	SHMM (2)	37.83	72.2
English	HSHMM	**40.04**	**76.60**
Mboshi	HMM	35.85	47.92
Mboshi	SHMM	38.38	**59.50**
Mboshi	SHMM (2)	36.09	53.06
Mboshi	HSHMM	**41.07**	59.15
Yoruba	HMM	36.38	54.47
Yoruba	SHMM	38.99	64.46
Yoruba	SHMM (2)	36.97	58.59
Yoruba	HSHMM	**40.06**	**66.95**

Table: Comparison of the HMM, SHMM and HSHMM

SHMM (2): the subspace is retrained on the target language.


Units frequency (timit)
Units frequency (mboshi)
Units frequency (yoruba)
reference
hmm
shmm
hshmm
vqvae
0
20
40
60
80
100
Units duration (timit)
Units duration (mboshi)
Units duration (yoruba)
0.00
0.05
0.10
0.15
0.20
0.25
time (s)



sil 78 11 97 39 22 63 28 41 7 1 70 39 79 100 75 2 43 22 4 36 28 65 95 8 95 28 65 30 74 44 49 87 80 sil
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
HMM
sil 33 50 30 54 52 49 6 38 76 30 75 15 37 11 83 8 51 44 56 23 56 26 57 40 77 45 46 7 78 74 sil
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
SHMM
sil 13 74 14 9 94 13 sil 27 67 9 94 28 66 2 27 2 30 28 44 7 44 30 27 11 7 67 sil
h# s ax m eh n tcl t ih z m eh zh axr dcl d ih n kcl k y ux bcl b ix kcl k y aa r dcl d s h#
HSHMM

Outline

Summary

- We have proposed two new models for the task of Acoustic Unit Discovery:
  - Subspace Hidden Markov Model
  - Hierarchical Subspace Hidden Markov Model
- These models are inspired by how infants learn to speak
- They show strong improvement in terms of clustering and segmentation quality
- The concept of (hierarchical) subspace and can be extended to a large class of models
- To reproduce our experiments:
  `https://github.com/beer-asr`

Future Challenges

- AUD is not a solved problem!
- Model suffers from high variability of speech
- Two major problems:
  - Acoustic modeling: going beyond HMM
  - Language modeling: discovery words
- Towards the first system to learn speech as humans...

Lucas Ondel, Hari Krishna Vydana, Lukáš Burget, and Jan Černocký (2019). "Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery". In: Proc. Interspeech 2019, pp. 261–265. URL: http://dx.doi.org/10.21437/Interspeech.2019-2224.

Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukáš Burget, Francois Yvon, and Sanjeev Khudanpur (2018). "Bayesian Models for Unit Discovery on a Very Low Resource Language". In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5939–5943. URL: https://ieeexplore.ieee.org/document/8461545.

Lucas Ondel, Lukaš Burget, Santosh Kesiraju, and Jan Černocký (2017). “Bayesian phonotactic language model for acoustic unit discovery”. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5750–5754. URL: https://www.fit.vut.cz/research/publication/11472/.en.

Lucas Ondel, Lukáš Burget and Jan Černocký (2016). “Variational inference for acoustic unit discovery”. In: Procedia Computer Science 81, pp. 80–86. URL: https://www.sciencedirect.com/science/article/pii/S1877050916300473.

Thank you for your attention.